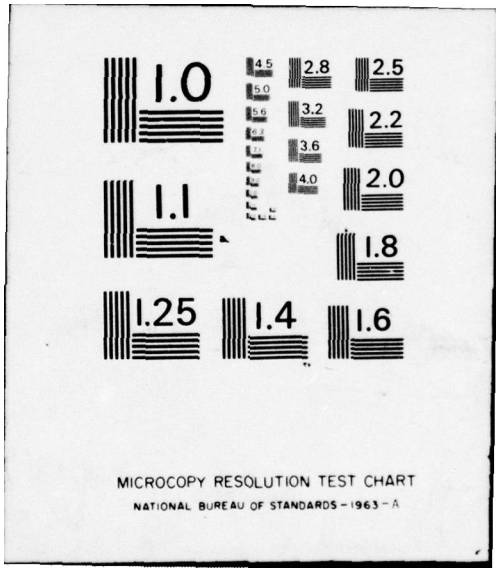
AD-A033 287 PORDUE UNIV LAFAYETTE IND DEPT OF STATISTICS F/G 12/1
CHI-SQUARE TESTS.(U)
1976 D S MOORE
UNCLASSIFIED MIMEOGRAPH SER-469 AFOSR-TR-76-1227 AF-AFOSR-2350-72 NL

END
DATE FILMED
2-77


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS - 1963 - A

AFOSR - TR - 76 - 1227

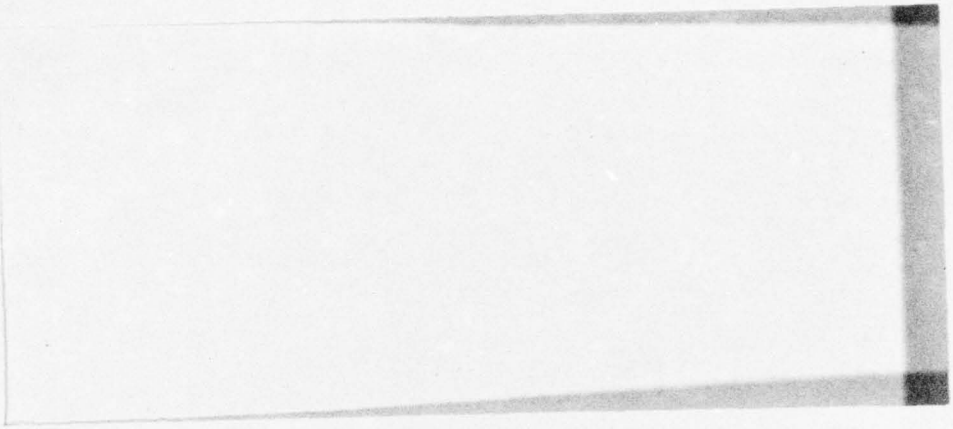
(3)

PURDUE UNIVERSITY

DEPARTMENT OF STATISTICS

DIVISION OF MATHEMATICAL SCIENCES

Approved for public release;
distribution unlimited.



CHI-SQUARE TESTS*

by

David S. Moore
Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #469

*To appear in R. V. Hogg (Ed.) Studies in Statistics, Mathematical Association of America. Preparation of this paper was supported by the Air Force Office of Scientific Research under Grant AFOSR-72-2350. The United States Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation hereon.

CHI-SQUARE TESTS*

David S. Moore
Purdue University

1. INTRODUCTION

Statistics is the science of collecting, describing and interpreting data. The most common mathematical model underlying statistical interpretation of data assumes that the values of measured variables in the population of interest are described by a probability distribution. If several variables are measured (say the length and weight of a cockroach), the population is described by a multivariate probability distribution. When the forces of good prevail, the fortunate statistician has data consisting of observed values of independent random variables, each having the population probability distribution. The statistical design of sampling and experimentation is intended to produce this happy state of affairs or some moderate complication of it. We will assume that, whether by design or (this is risky) by good fortune, the data collection process yields independent random variables $X_1,\ldots,X_n$ having a common probability distribution. This distribution is unknown - that's the distinction between statistics and probability theory. Let F denote the unknown distribution function (df) of any single X_j.

It is clear that a classical statistical problem is "Which probability models adequately describe the data?" This question can be asked for descriptive purposes or as a preliminary to formal inference from the data.

*Preparation of this paper was supported by the Air Force Office of Scientific Research under Grant AFOSR-72-2350.

Particularly in the latter case, the statistician may have in mind a specific family of probability distributions (such as the normal family) and the more exact question "Do the data support or impugn the hypothesis that the population distribution is a member of this family?" Most common families of distributions have df's of specified functional form indexed by a (real or vector) parameter. For example, an individual member of the univariate normal family of distributions is specified by the values of the mean μ and standard deviation σ. If $G(\cdot|\theta)$ is a family of df's indexed by a parameter θ running over a parameter space Ω, we have now formulated the following problem.

Given independent random variables $X_1,\ldots,X_n$ *having common unknown* df F, *test the hypothesis*

$$H_0: \quad F(\cdot) = G(\cdot|\theta) \text{ for some } \theta \text{ in } \Omega$$

This is the problem of *goodness of fit*. Notice that in practice the observations X_j will often be multivariate, and that the null hypothesis will usually be composite (that is, the family $G(\cdot|\theta)$ will contain more than one member). Notice also that although we have stated the problem in terms of hypothesis testing, it will rarely be sensible to simply accept or reject at the usual significance levels such as $\alpha = .05$. In particular, if we test fit to (say) the univariate normal family as a preliminary to a further analysis which assumes normality, we should surely not cling to the assumption of normality until the evidence against it is significant at the five percent level. Many applied statisticians favor using an α of .20 or .25 for such preliminary tests. The real difficulty is that the H_0 in the problem of fit does not have the status ("The statement we hope to find evidence against") ascribed to null hypotheses in standard tests of significance. Nonetheless, the attained significance level of a test of fit is at least a descriptive measure of the distance of the data

from the hypothesized family of distributions. We will therefore study the theory of some tests of fit without further ventures into the wilderness of philosophies of inference.

The oldest family of tests of fit was fathered by Karl Pearson in 1900. During the preceding decade, Pearson had developed families of probability distributions in the course of his work on *Mathematical Contributions to the Theory of Evolution*. He now wished to see which of these fit his data, rather than simply assuming that all biological variables are normally distributed. Statistics as a discipline was in its infancy in 1900. Many results and methods which would form part of the new science were scattered through the work of Gauss, Laplace, Lagrange and others, but these results were not collected and unified, and were often unknown to statisticians such as Pearson. The binomial distributions and their approximation by normal distributions were well known; the chi-square distributions were known as the distributions of sums of squares of independent normal random variables; and the multivariate normal distributions had only recently become familiar. These last distributions will play a major role in our study. Pearson knew the p-variate normal distribution with mean vector μ and nonsingular covariance matrix Σ as the distribution having density function of the form

$$(1) \qquad f(y') = ce^{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)}$$

Here $y' = (y_1,\ldots,y_p)$ is the p-variate argument of the density function. If $Y = (Y_1,\ldots,Y_p)'$ is a random variable having this distribution, we will write $Y \sim N_p(\mu,\Sigma)$ to express this fact.

Pearson sought first to test the simple null hypothesis that univariate observations $X_1,\ldots,X_n$ have a given df G. He partitioned the line into cells $E_1,\ldots,E_M$ and based his test on the frequencies $N_1,\ldots,N_M$ of observations falling in these cells. If the hypothesis is true and

$$p_i = P_G[X \text{ in } E_i] = \int_{E_i} dG(x),$$

then np_i is the expected frequency for E_i and the quantities $N_i - np_i$ measure the lack of fit between data and model. Pearson then argued:

(a) If $Y \sim N_p(0,\Sigma)$, then the quadratic form which appears in the exponent of the density function (1) has the same distribution as the sum of squares of p independent standard normal random variables. This is the chi-square distribution with p degrees of freedom, and we write $Y'\Sigma^{-1}Y \sim \chi^2(p)$.

(b) By the DeMoivre-Laplace normal approximation to the binomial distributions, each N_i-np_i is approximately normal when the number of observations n is large.

(c) Computing variances and covariances for $Y = (N_1 - np_1,\ldots,N_{M-1} - np_{M-1})'$ and inverting the covariance matrix shows that

$$Y'\Sigma^{-1}Y = \sum_{i=1}^{M} \frac{(N_i - np_i)^2}{np_i},$$

and this statistic therefore has approximately the $\chi^2(M-1)$ distribution when the null hypothesis is true and n is large. Large values of this statistic (i.e., values in the upper tail of the $\chi^2(M-1)$ distribution) are evidence of lack of fit.

This argument contains some minor mathematical gaps: it ignores the distinction between approximate normality of each $N_i - np_i$ and approximate multivariate normality of the vector Y, and does not show how (a) implies that when Y is *approximately* $N_{M-1}(0,\Sigma)$, then $Y'\Sigma^{-1}Y$ is approximately $\chi^2(M-1)$.

But the argument is in the best spirit of pre-Weierstrass mathematics, needing only a few technicalities to become rigorous. Pearson's proof shows, for example, why his famous chi-square statistic does not have the variance $np_i(1-p_i)$ of N_i in the denominator of the ith summand. More important, the idea of reducing the general problem of fit to a combinatorial setting (counting numbers of observations in each of M cells) was of lasting significance. Chi-square tests remain among the most common tests of fit, largely because of the flexibility of Pearson's idea. If, for example, the observations X_j and the cells E_i are multidimensional, the distribution of the cell frequencies N_i and the form and theory of the Pearson chi-square statistic are unchanged.

Now of course the null hypothesis in a problem of fit is generally that the df of the observations falls in a family $\{G(\cdot|\theta): \theta \text{ in } \Omega\}$ of df's. In this case, the cell probabilities depend on the unknown parameter θ,

$$p_i(\theta) = \int_{E_i} dG(x|\theta).$$

Pearson proposed to estimate the unknown parameter from the data by some function $\theta_n = \theta_n(X_1,\ldots,X_n)$. In testing fit to the univariate normal family, for example, the parameter is $\theta = (\mu,\sigma)$ and the population mean μ and standard deviation σ are commonly estimated by the mean and standard deviation of the sample. The Pearson statistic now becomes

$$\sum_{i=1}^{M} \frac{[N_i - np_i(\theta_n)]^2}{np_i(\theta_n)} \quad . \tag{2}$$

That is, we test the fit of the data to the df $G(\cdot|\theta_n)$ having the estimated parameter value.

Unfortunately, as mathematicians learned before statisticians, pre-Weierstrass mathematics has its limitations. Even when the estimator θ_n approaches the true value of θ as the sample size n increases, the statistic (2) does *not* then have approximately the $\chi^2(M-1)$ distribution, as Pearson believed it did. Since statistical methods are actually used in the real world, observant users began to suspect that something was amiss. Some even did extensive simulations (quite a chore in those pre-computer days) to compare Pearson's theoretical distribution with the observed distribution of his statistic. It was not until 1924 that R. A. Fisher showed that the statistic (2) does not have approximately the $\chi^2(M-1)$ distribution in large samples, and that the distribution it does have depends on how the unknown parameter is estimated. If θ_n is the value of θ which minimizes the statistic (2) for given N_i (so that $G(\cdot|\theta_n)$ is the closest df in the hypothesized family to the data by this measure of distance), Fisher showed that the approximate distribution is $\chi^2(M-m-1)$ when θ is m-dimensional. When other methods of estimation are used, this conclusion is false. It is false, for example, when the sample mean and standard deviation are used in testing fit to the univariate normal family. Only since the 1950's has a rigorous study of chi-square statistics with general estimators θ_n been made, and solutions obtained to many practical and mathematical problems concerning these statistics.

Our study of the modern theory of chi-square tests of fit will touch on several other topics in statistics which are important in their own right. We must first acquaint ourselves with the multivariate normal family of distributions. And since chi-square tests are large-sample tests, based on the limiting multivariate normal distribution of the cell frequencies, some of the basic techniques of statistical large sample theory must be mentioned. Finally, Pearson's construction of the proper quadratic form

in the cell frequencies is the genesis of some familiar statistical procedures other than tests of fit. In all of this, we hope not to entirely lose sight of the interplay between theory and practice which gives statistics its vitality.

2. THE MULTIVARIATE NORMAL DISTRIBUTIONS

The multivariate normal family plays a role in the study of multi-dimensional data analogous to that played by the univariate normal family in one dimension. These distributions are not only important probability models in their own right, but because of the central limit theorem serve as large sample approximations to other models. We will *not* define these distributions by the density function (1) for two reasons. First, that definition is awkward due to the complexity of the density function. More important, a distribution in Euclidean p-space R^p does not have a density function in R^p if it assigns probability one to a set of measure zero. Such *singular distributions* - other than the discrete distributions supported on a countable set - are somewhat pathological in one dimension, and play little role in probability modeling there. But in higher dimensions it is quite common for random variables to be dependent in such a way that with probability 1 their values fall in a lower-dimensional hyperplane and their joint distribution is thus singular. The cell frequencies $N_1,\ldots,N_M$ in a chi-square test, for example, satisfy $\Sigma_{i=1}^{M} N_i = n$ and so take values only in this (M-1)-dimensional hyperplane. In Section 1 we followed Pearson in working with the nonsingular distribution of $(N_1,\ldots,N_{M-1})'$, but this mode of escape is more awkward in other settings.

To statisticians, the most useful definition of a probability distribution is often a *representational definition* - a statement of what random variable has the distribution. For example, the $\chi^2(p)$ distribution is that of $\Sigma_{i=1}^{p} Z_i^2$ where $Z_1,\ldots,Z_p$ are independent $N(0,1)$ random variables. In this spirit, *the* $N_p(\mu,\Sigma)$ *distribution is defined as the distribution of the random variable*

$$Y = AZ + \mu \tag{3}$$

where $Z = (Z_1,\ldots,Z_m)'$ *and* $Z_i \sim N(0,1)$ *and are independent,* $\mu = (\mu_1,\ldots,\mu_p)'$ *is the vector of means, and* A *is any* $p\times m$ *matrix satisfying* $AA' = \Sigma$. That is, the multivariate normal distributions are the distributions of affine transformations of a set of independent standard normal random variables.

It is easy to check that μ and $AA' = \Sigma$ are in fact the mean vector and covariance matrix of the p-variate random vector defined in (3). To justify this definition of $N_p(\mu,\Sigma)$, we must show that Y so defined has the <u>same</u> distribution for any m and any $p\times m$ matrix A satisfying $AA' = \Sigma$. To justify the notation $N_p(\mu,\Sigma)$, we must show that this family is parameterized by (μ,Σ) alone. Both of these facts follow from a computation of the characteristic function (Fourier transform) of the distribution of Y. This computation also illustrates the convenience of the representational definition.

The characteristic function of Y is the function of p real variables $t' = (t_1,\ldots,t_p)$ defined by

$$\begin{aligned}\varphi_Y(t') &= E[e^{it'Y}] \\ &= E[e^{i(t'AZ+t'\mu}] \\ &= e^{it'\mu}E[e^{it'AZ}]\end{aligned}$$

Now since the characteristic function of any $Z_j \sim N(0,1)$ is easily computed to be

$$E[e^{isZ_j}] = e^{-\frac{1}{2}s^2} \qquad -\infty < s < \infty$$

and $Z_1,\ldots,Z_m$ are independent,

$$\begin{aligned}\varphi_Y(t') &= e^{it'\mu}E[e^{i\{(t'A)_1Z_1+\ldots+(t'A)_mZ_m\}}] \\ &= e^{it'\mu}\prod_{j=1}^{m} E[e^{i(t'A)_jZ_j}] \\ &= e^{it'\mu}\prod_{j=1}^{m} e^{-\frac{1}{2}[(t'A)_j]^2} \\ &= e^{it'\mu-\frac{1}{2}t'AA't} = e^{it'\mu-\frac{1}{2}t'\Sigma t}.\end{aligned}$$

This characteristic function is the same for all m and A with $AA' = \Sigma$, and is parameterized by μ and Σ. Since the characteristic function uniquely determines a probability distribution, $N_p(\mu,\Sigma)$ is well defined.

The definition (3) is geometrically transparent. The distribution of Z is nonsingular in R^m - indeed, has all of R^m as its support. If the linear transformation A: $R^m \to R^p$ has full rank p, then the distribution $N_p(\mu,\Sigma)$ of Y is nonsingular and has R^p as its support. If the rank of A is $r < p$, then $N_p(\mu,\Sigma)$ is singular and is supported on the r-dimensional hyperplane in R^p obtained by translating the range of A by μ. Since the range of $AA' = \Sigma$ is the same as the range of A, $N_p(\mu,\Sigma)$ *is a nonsingular distribution if and only if the covariance matrix* Σ *is nonsingular. The support of* $N_p(\mu,\Sigma)$ *is a hyperplane in* R^p *with dimension equal to the rank of* Σ.

Many properties of the multivariate normal distributions follow easily from (3). For example, if B is any s×p matrix, then applying B to both sides of (3) shows at once that $BY \sim N_s(B\mu, B\Sigma B')$. That is, *any set of linear combinations of jointly normal random variables has again a multivariate normal joint distribution.* In particular, any single linear combination is univariate normal, and this includes each individual component Y_i of Y.

That the random variable Y has density function of the form (1) when Σ is nonsingular can be deduced by a change of variables in the known density function of Z.

The joint distribution of a set of independent univariate normal random variables is a multivariate normal distribution with a diagonal covariance matrix Σ. (Because $N(\mu_i,\sigma_i^2)$ is the distribution of $Y_i = \sigma_i Z_i + \mu_i$ for $Z_i \sim N(0,1)$, so that (3) fits $Y = (Y_1,\ldots,Y_p)'$.) Since $N_p(\mu,\Sigma)$ is determined by μ and Σ, it follows that *random variables having a multivariate normal joint distribution are independent if and only if they are uncorrelated.* Independence in a multivariate normal setting can therefore be established by simply computing covariances. If I_p denotes the p×p identity matrix, $N_p(0,I_p)$ now denotes the distribution of a set of p independent N(0,1) random variables. If $Z \sim N_p(0,I_p)$ and P is a p×p orthogonal matrix, it follows that $PZ \sim N_p(0,I_p)$ once again.

For our study of chi-squared tests, we are particularly interested in quadratic forms in multivariate normal random variables. The representational approach reduces this to the study of quadratic forms in independent N(0,1) random variables. We know one fact about such forms: if $Z \sim N_p(0,I_p)$ then the sum of squares $Z'Z = \Sigma_{i=1}^p Z_i^2$ has the $\chi^2(p)$ distribution. Two potential generalizations of this fact come to mind at once. When $Y \sim N_p(0,\Sigma)$ we can ask (1) What quadratic forms Y'CY have chi-square distributions? (2) What is the distribution of the sum of squares Y'Y? Since partial sums of squares in the $N_p(0,I_p)$ case have $\chi^2(k)$ distributions for $k < p$, the first question might be specialized to: What quadratic forms Y'CY have the $\chi^2(r)$ distribution for r as large as possible?

It is convenient for the study of quadratic forms to use a particular representation of $Y \sim N_p(0,\Sigma)$ based on the fact that *any nonnegative definite symmetric matrix (such as an arbitrary covariance matrix Σ) has a unique nonnegative definite symmetric square root.* To see this, note first that any square root $\Sigma^{\frac{1}{2}}$ commutes with its square Σ, so that if $\Sigma^{\frac{1}{2}}$ is symmetric, $\Sigma^{\frac{1}{2}}$ and Σ can be simultaneously diagonalized by some orthogonal matrix P. From

$$(4) \qquad P\Sigma P' = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_p \end{pmatrix} = D$$

$$P\Sigma^{\frac{1}{2}}P' = \begin{pmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_p \end{pmatrix}$$

and $(\Sigma^{\frac{1}{2}})^2 = \Sigma$ it follows that $\mu_i = \sigma_i^{\frac{1}{2}}$ and hence that all symmetric nonnegative definite square roots have the form

$$\Sigma^{\frac{1}{2}} = P'\begin{pmatrix} \sigma_1^{\frac{1}{2}} & & \\ & \ddots & \\ & & \sigma_p^{\frac{1}{2}} \end{pmatrix}P$$

for P and D satisfying (4) and nonnegative square roots $\sigma_i^{\frac{1}{2}}$. This also shows that such a $\Sigma^{\frac{1}{2}}$ exists. The P and D in (4) are not unique, but are determined up to permutations of the σ_i and of the corresponding rows of P. It is easy to see that $\Sigma^{\frac{1}{2}}$ is unchanged by such permutations and is hence unique.

We can thus represent $Y \sim N_p(0,\Sigma)$ as $Y = \Sigma^{\frac{1}{2}}Z$. The distribution of quadratic forms in Y is completely described by the following result.

THEOREM 1. *Suppose that* $Y \sim N_p(0,\Sigma)$ *and that* C *is any* p×p *symmetric matrix. Then the quadratic form* Y'CY *has the distribution of* $\Sigma_{i=1}^{p} \lambda_i Z_i^2$, *where the* Z_i *are independent* N(0,1) *random variables and the* λ_i *are the characteristic roots of* $\Sigma^{\frac{1}{2}}C\Sigma^{\frac{1}{2}}$.

PROOF. Since $Y = \Sigma^{\frac{1}{2}}Z$,

$$Y'CY = Z'\Sigma^{\frac{1}{2}}C\Sigma^{\frac{1}{2}}Z = Z'QZ.$$

Because $Q = \Sigma^{\frac{1}{2}}C\Sigma^{\frac{1}{2}}$ is symmetric, there is an orthogonal matrix P such that PQP' = D, where D is diagonal with the λ_i as diagonal elements. So Q = P'DP and

$$Y'CY = Z'QZ = (PZ)'D(PZ)$$

The right side above is $\Sigma_{i=1}^{p} \lambda_i (Z_i^*)^2$ where Z_i^* is the ith component of PZ. Since $PZ \sim N_p(0,I_p)$, this is a representation of the desired form.

When $X \sim N_p(\mu,\Sigma)$ and Σ is nonsingular, $Y = X - \mu \sim N_p(0,\Sigma)$ and we obtain as a corollary of Theorem 1 Pearson's result that the quadratic form $(X-\mu)'\Sigma^{-1}(X-\mu)$ appearing in the exponent of the density function has the distribution of $\Sigma_{i=1}^{p} Z_i^2$, which is $\chi^2(p)$. This answers our first question when Σ is nonsingular. The answer to the second question, concerning the distribution of the sum of squares Y'Y, is answered by setting $C = I_p$ in Theorem 1. The distribution is that of $\Sigma_{i=1}^{p} \lambda_i Z_i^2$, where the λ_i are now the characteristic roots of Σ itself.

We have yet to answer the first question fully by extending Pearson's recipe to the singular case. Since the rank of $\Sigma^{\frac{1}{2}}C\Sigma^{\frac{1}{2}}$ cannot exceed the rank (say r) of Σ and $\Sigma^{\frac{1}{2}}$, it is clear that if $YCY \sim \chi^2(k)$, then $k \leq r$. Theorem 1 implies that $Y'CY \sim \chi^2(k)$ if and only if $\Sigma^{\frac{1}{2}}C\Sigma^{\frac{1}{2}}$ is idempotent of rank k, and a little matrix manipulation shows that a sufficient condition for this is

that ΣC be idempotent of rank k. This general result is not very helpful in the search for C such that $Y'CY \sim \chi^2(r)$. Pearson's result involved the inverse Σ^{-1}. It turns out that an "inverse" for singular matrices neatly generalizes his recipe.

For an arbitrary n×m real matrix A, a *generalized inverse* of A can be defined by any of the following statements.

(a) A generalized inverse of A is any m×n matrix G such that $x = Gy$ solves the equations $y = Ax$ for any y in the range (column space) of A.

(b) A generalized inverse of A is any m×n matrix G satisfying $AGA = A$.

(c) A generalized inverse of A is any m×n matrix G such that AG is a projection onto the range of A.

It is not hard to show that these definitions are equivalent. Definition (a) justifies the concept in terms of solving consistent sets of linear equations with matrix A. Definition (b) is convenient for matrix manipulation, while (c) gives some geometric insight. Note that in (c) "projection" does not mean the unique orthogonal projection onto the range of A, but any idempotent matrix with this range. When A is not a nonsingular square matrix, it possesses many generalized inverses. Any generalized inverse of A will be denoted by A^-. Generalized inverses are widely used to provide a unified notation for linear statistical problems when matrices may be singular. The following theorem and its proof illustrate the convenience of this notion.

THEOREM 2. *Suppose that* $Y \sim N_p(0,\Sigma)$ *and* Σ *has rank* r. *Then*

(a) *With probability* 1, $Y'\Sigma^-Y$ *is the same for all choices of* Σ^-.

(b) $Y'\Sigma^-Y \sim \chi^2(r)$ *and is the unique quadratic form having this distribution.*

PROOF. (a) If x is any vector in the range of Σ, so that $x = \Sigma y$ for some y, then

$$x'\Sigma^-x = y'\Sigma\Sigma^-\Sigma y = y'\Sigma y$$

by definition (b) and symmetry of Σ. So $x'\Sigma^-x$ is the same number for all choices of Σ^- whenever x is in the range of Σ. But $Y = \Sigma^{\frac{1}{2}}Z$ is in the range of Σ(which is the same as that of $\Sigma^{\frac{1}{2}}$) with probability 1.

(b) Since $Y'\Sigma^-Y$ is the same for all choices of Σ^-, we can chose a convenient generalized inverse. If Σ has rank r and positive characteristic roots $d_1,\ldots,d_r$, there is an orthogonal P such that

$$P\Sigma P' = \begin{pmatrix} d_1 & & & & \\ & \ddots & & & \\ & & d_r & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix}$$

An obvious choice of generalized inverse is

$$\Sigma^- = P'\begin{pmatrix} d_1^{-1} & & & & \\ & \ddots & & & \\ & & d_r^{-1} & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix} P.$$

Since

$$\Sigma^{\frac{1}{2}} = P'\begin{pmatrix} d_1^{\frac{1}{2}} & & & & \\ & \ddots & & & \\ & & d_r^{\frac{1}{2}} & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix} P$$

we obtain

$$Y'\Sigma^{-}Y = Z'\Sigma^{\frac{1}{2}}\Sigma^{-}\Sigma^{\frac{1}{2}}Z = \sum_{i=1}^{r}(Z_i^*)^2$$

where $Z^* = PZ \sim N_p(0,I_p)$. Thus $Y'\Sigma^{-}Y \sim \chi^2(r)$.

It remains to show that $Y'CY \sim \chi^2(r)$ implies that C is a generalized inverse of Σ. By Theorem 1, $Y'CY \sim \chi^2(r)$ if and only if $\Sigma^{\frac{1}{2}}C\Sigma^{\frac{1}{2}}$ is idempotent of rank r. But then $\Sigma^{\frac{1}{2}}C\Sigma^{\frac{1}{2}}$ is a projection, and since its range is contained in the range of $\Sigma^{\frac{1}{2}}$ and has the same dimension r, it is a projection onto the range of $\Sigma^{\frac{1}{2}}$. A projection acts as the identity transformation on its range, so

$$\Sigma C\Sigma = \Sigma^{\frac{1}{2}}(\Sigma^{\frac{1}{2}}C\Sigma^{\frac{1}{2}})\Sigma^{\frac{1}{2}} = \Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}} = \Sigma$$

and C satisfies definition (b) of Σ^{-}.

When $Y \sim N_p(\mu,\Sigma)$, the representation $Y = \Sigma^{\frac{1}{2}}Z + \mu$ can be applied to the study of quadratic forms in Y. Repeating the argument of Theorem 1 shows that Y'CY has the distribution of a random variable of the form

$$\sum_{i=1}^{p}\lambda_i Z_i^2 + 2\sum_{i=1}^{p} b_i Z_i + c \tag{5}$$

where $Z \sim N_p(0,I_p)$ and the λ_i are as in Theorem 1. These distributions have no neat classification. We will make only one foray into this "noncentral case," to look again at Pearson's recipe.

When $Z \sim N_p(0,I_p)$, $Z'Z \sim \chi^2(p)$ by definition. When $Y \sim N_p(\mu,I_p)$, the distribution of Y'Y, or equivalently of $(Z+\mu)'(Z+\mu)$, is defined to be the *noncentral chi-square distribution with* p *degrees of freedom and noncentrality parameter* $\delta = \mu'\mu$. (Since the statistic is the square of the distance of Z from the point $-\mu$ in R^p, it follows from the circular symmetry of the density function of Z that this distribution depends only on the distance of $-\mu$ from the origin. Thus parameterizing the distribution by (p,δ) is

justified.) We will use the notation $Y'Y \sim \chi^2(p,\delta)$.

Suppose now that $Y \sim N_p(\mu,\Sigma)$ and Σ has rank r. What then is the distribution of the generalized Pearson statistic $Y'\Sigma^-Y$? Alas, since $Y = \Sigma^{\frac{1}{2}}Z + \mu$, Y is not in the range of Σ unless μ is, so that the quadratic form $Y'\Sigma^-Y$ changes with the choice of Σ^-. If μ is in the range of Σ, we can write $\mu = \Sigma^{\frac{1}{2}}\nu$ and follow the argument of Theorem 2 to show that $Y'\Sigma^-Y$ is well defined and that

$$Y'\Sigma^-Y = (Z+\nu)'\Sigma^{\frac{1}{2}}\Sigma^-\Sigma^{\frac{1}{2}}(Z+\nu)$$

$$= \sum_{i=1}^{r} (Z_i + \nu_i)^2 \sim \chi^2(r,\delta)$$

where $\delta = \Sigma_{i=1}^{r} \nu_i^2$. But by the same argument,

$$\sum_{i=1}^{r} \nu_i^2 = \nu'\Sigma^{\frac{1}{2}}\Sigma^-\Sigma^{\frac{1}{2}}\nu = \mu'\Sigma^-\mu.$$

Thus $Y'\Sigma^-Y \sim \chi^2(r,\mu'\Sigma^-\mu)$. When μ is not in the range of Σ, both the form and the distribution of $Y'\Sigma^-Y$ vary with the choice of Σ^-. Of course, when Σ is nonsingular these complications do not arise, and $Y'\Sigma^{-1}Y \sim \chi^2(p,\mu'\Sigma^{-1}\mu)$ for any mean vector μ.

We have concentrated on cases in which the distribution of Y'CY given by Theorem 1 (or more generally by (5)) reduces to a chi-square distribution. There are sound practical reasons for doing so, even though machine computation makes it feasible to produce tables of critical points for the distributions of $\Sigma_{i=1}^{p} \lambda_i Z_i^2$. Tests of fit based on quadratic forms in (approximately) multivariate normal random variables are the natural generalization of Pearson's chi-square test. These tests must compete for the attention of practical statisticians against special-purpose tests for fit to specific common families, and against general tests of fit based on

the empiric distribution function (EDF tests). These competitors are usually much more powerful than chi-square tests, but are also much less flexible in adapting to unknown parameters and discrete or multivariate data. In particular, they require separate computation of critical points for each hypothesized family. (I will not mention - this is a rhetorical device I learned from Cicero - that the EDF tests break down almost completely when faced with hypothesized distributions which are multi-dimensional or are not location-scale families.) If a test of chi-square type also requires a special computation of critical points to be applicable to a given problem, we would usually be wiser to allot our computer time to an EDF test instead. Thus generalizations of Pearson's statistic lose much of their attractiveness if their critical points cannot be found in standard tables. In the light of Theorem 1, the relevant tables will be those for the chi-square distributions.

3. LARGE SAMPLE THEORY

Since the earliest days of statistics it has been noticed that complicated distributions often have simple approximations for large samples. The distribution of Pearson's chi-square statistic is an example. The use of chi-square tests, both as tests of fit and for other common applications, is based on approximating multinomial distributions by the multivariate normal distributions which are their limits as the sample size increases. We will therefore review some facts about statistical large sample theory. There are three major aspects to this theory. The first simply asks questions of convergence: "What happens in the limit?" The second studies the approach to the limit by providing rates of convergence, asymptotic expansions, etc. The third considers the usefulness of the

asymptotic forms provided by the first two parts of the subject as approximations to the fixed sample size truth. Explicit numerical calculation and simulation play large roles here. Only the first aspect of large sample theory will concern us, both for simplicity's sake and because (to make an appalling generalization) in the field of chi-square tests the second aspect has had little practical impact and the third has shown that use of limiting distributions is an adequate approximation for quite moderate sample sizes.

The most useful mode of convergence for statistical use is convergence in distribution. If $X_1, X_2, \ldots$ are R^p-valued random variables, X_n having df F_n, we say that the sequence *converges in distribution* to the distribution having df F if $F_n(x) \to F(x)$ for every continuity point x of F. Abusing notation to also denote by F, F_n the probability measures on R^p generated by these df's, convergence in distribution is equivalent to

$$\lim_n P[X_n \text{ in } A] = \lim_n F_n(A) = F(A)$$

for all Borel sets A in R^p whose boundaries have probability zero under F. Thus $P[X_n \text{ in } A]$ can be approximated by F(A) for large n. Convergence in distribution to the distribution placing probability 1 on a single point c is equivalent to *convergence in probability* of X_n to c. That is, for any $\varepsilon > 0$, $P[|X_n - c| > \varepsilon] \to 0$ as $n \to \infty$. (We write this $X_n \to c(P)$.) All of this is of course a province of the measure theory which underlies statistical theory and sometimes invades the conscious thought of the working statistician. A nice exposition in effortless generality appears in the first chapter of [1].

We require only one specific and two general facts about convergence in distribution. The specific fact is *the multivariate central limit theorem: If* $X_1, X_2, \ldots$ *are independent* R^p*-valued random variables having a common distribution with vector of means* μ *and finite* $p \times p$ *covariance matrix* Σ*, and if* $\bar{X}_n = n^{-1}\Sigma_1^n X_i$*, then* $n^{\frac{1}{2}}(\bar{X}_n - \mu)$ *converges in distribution to* $N_p(0,\Sigma)$. This is written

$$n^{\frac{1}{2}}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} N_p(0,\Sigma).$$

We often abuse notation and write instead

$$n^{\frac{1}{2}}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} Y$$

where $Y \sim N_p(0,\Sigma)$, even though convergence in distribution makes no statement about convergence of values of $n^{\frac{1}{2}}(\bar{X}_n - \mu)$ or about any limiting random variable.

The essential general fact is *the continuity theorem: If* $Y_n \xrightarrow{\mathcal{D}} Y$ *and* $h: R^p \to R^k$ *is continuous with probability* 1 *with respect to the distribution of* Y*, then* $h(Y_n) \xrightarrow{\mathcal{D}} h(Y)$. The continuity theorem licenses our natural desire to conclude that when (say) Y_n is approximately $N_p(0,\Sigma)$, then $Y_n'CY_n$ has approximately the distribution specified by Theorem 1. The central limit theorem provides us with a large supply of random variables which are approximately multivariate normal. The two together suffice to make rigorous Pearson's proof outlined in Section 1 above.

The second general fact is needed to still the clamoring voices of the pedants. *If* $X_n \xrightarrow{\mathcal{D}} X$ *and* $Y_n \to c(P)$*, then* $(X_n,Y_n) \xrightarrow{\mathcal{D}} (X,c)$. That is, convergence of both marginal distributions of (X_n,Y_n) suffices for convergence of the joint distribution *if* one sequence of marginal distributions has a degenerate

limit. Convergence of marginal distributions in general gives no information about the joint distribution. The natural manipulations we wish to make are all licensed by these two general facts. For example, if $X_n \xrightarrow{\mathcal{D}} X$ and $R_n \to 0(P)$, then $X_n + R_n \xrightarrow{\mathcal{D}} X$, as reason and justice demand. For $(X_n, R_n) \xrightarrow{\mathcal{D}} (X,0)$ by the second fact, and the continuity theorem now applies with $h(x,y) = x+y$.

The following section will provide examples in plenty of the way in which the three facts mentioned here combine with the law of large numbers and Taylor's theorem to form the elementary arithmetic of statistical large sample theory.

4. CHI-SQUARE TESTS OF FIT

Returning at last to the problem and notation of Section 1, we wish to test whether independent random variables $X_1,\ldots,X_n$ taking values in Euclidean p-space R^p have df $G(\cdot|\theta)$ for some θ in Ω, an open set in R^m. Partitioning R^p into M cells $E_1,\ldots,E_M$, we denote by N_i the number of $X_1,\ldots,X_n$ falling in E_i and by $p_i(\theta)$ the probability that a random variable with df $G(\cdot|\theta)$ falls in E_i. The vector of standardized cell frequencies is the M-vector $V_n(\theta)$ with ith component

$$\frac{N_i - np_i(\theta)}{[np_i(\theta)]^{\frac{1}{2}}}.$$

Finally, θ is estimated from $X_1,\ldots,X_n$ by $\theta_n = \theta_n(X_1,\ldots,X_n)$, and $C_n = C_n(X_1,\ldots,X_n)$ is a possibly data-dependent non-negative definite symmetric M×M matrix. *Statistics of chi-square type are statistics of the form*

$$V_n(\theta_n)'C_nV_n(\theta_n), \tag{6}$$

that is, non-negative definite quadratic forms in the standardized cell frequencies.

If the vector $V_n(\theta_n)$ has a limiting $N_M(0,\Sigma(\theta_0))$ distribution when $G(\cdot|\theta_0)$ is the true df, and if $C_n \to C(\theta_0)(P)$, then the continuity theorem tells us that the limiting distributions of statistics of chi-square type under the null hypothesis are completely described by Theorem 1. Establishing asymptotic normality of $V_n(\theta_n)$ is therefore the primary mathematical hurdle in the theory of chi-square statistics. When θ, or more precisely the vector $p(\theta) = (p_1(\theta),\ldots,p_M(\theta))'$, is known, this hurdle is low indeed. For the N_i have a multinomial distribution with parameters n and $p(\theta)$. The vector $(N_1,\ldots,N_M)$ can be expressed as the sum of n independent M-dimensional indicator variables $\delta_1,\ldots,\delta_n$ where δ_j has ith component 1 and all others 0 when X_j falls in cell E_i. It follows from a computation of covariances and the multivariate central limit theorem that under $G(\cdot|\theta)$

$$V_n(\theta) \xrightarrow{\mathscr{D}} N_M(0, I_M - q(\theta)q(\theta)'), \tag{7}$$

where I_M is the M×M identity matrix and

$$q(\theta) = (p_1(\theta)^{\frac{1}{2}},\ldots,p_M(\theta)^{\frac{1}{2}})'.$$

This is just the multivariate normal approximation to a multinomial distribution, expressed in a notation which will prove convenient for easy extension to the more common case when θ must be estimated.

In that latter case, the asymptotic behavior of $V_n(\theta_n)$ will depend on that of θ_n, as Fisher recognized. Thus the large sample theory of chi-square statistics draws on the large sample theory of estimators, a main current of statistical theory since Fisher's time. Because of the importance of this subject, and to illustrate the application of the principles stated in Section 3, there follows an account of the large sample behavior of the

minimum chi-square estimator used in the classical Pearson test.

For given $N_1,\ldots,N_M$ this estimator is any value of θ which minimizes the Pearson statistic

$$P_n(\theta) = V_n(\theta)'V_n(\theta) = \sum_{i=1}^{M} \frac{[N_i - np_i(\theta)]^2}{np_i(\theta)} .$$

It is intuitively clear, and not hard to prove, that for such purposes as studying $V_n(\theta_n)$, the minimum chi-square estimator is asymptotically equivalent to the *minimum modified chi-square estimator* $\bar{\theta}_n$ which minimizes the modified chi-square statistic

$$Q_n(\theta) = \sum_{i=1}^{M} \frac{[N_i - np_i(\theta)]^2}{N_i} .$$

Working with $Q_n(\theta)$ is arithmetically simpler and conceptually identical to working with $P_n(\theta)$. We will therefore study $\bar{\theta}_n$, which we assume to exist and be a measureable function of $N_1,\ldots,N_M$. The first question concerns *consistency* of this estimator - does $\bar{\theta}_n$ approach the true value of θ as n increases?

LEMMA 1. *Suppose that* $M \geq m$, *that each* $\frac{\partial p_i}{\partial \theta_k}$ *is continuous at* $\theta = \theta_0$, *and that the* $M \times m$ *matrix* $\frac{\partial p_i}{\partial \theta_k}(\theta_0)$ *has rank* m. *Then any minimum modified chi-square estimator* $\bar{\theta}_n$ *satisfies* $\bar{\theta}_n \to \theta_0(P)$ *when* $G(\cdot|\theta_0)$ *is the true df of the* X_j.

PROOF. By the law of large numbers,

$$N_i/n \to p_i(\theta_0)(P) \qquad i = 1,\ldots,M \tag{8}$$

and therefore by the continuity theorem

$$Q_n(\theta_0)/n = \sum_{i=1}^{M} \frac{[N_i/n - p_i(\theta_0)]^2}{N_i/n} \to 0(P).$$

But by the definition of $\hat{\theta}_n$,

$$0 \leq Q_n(\bar{\theta}_n)/n \leq Q_n(\theta_0)/n$$

and so $Q_n(\bar{\theta}_n)/n \to 0(P)$. This can only happen if $N_i/n - p_i(\bar{\theta}_n) \to 0(P)$ for $i = 1,\ldots,M$. This with (8) implies that $p(\bar{\theta}_n) \to p(\theta_0)(P)$, which in turn implies that $\bar{\theta}_n \to \theta_0(P)$ if the function $\theta \to p(\theta)$ from R^m to R^M has a continuous inverse at $\theta = \theta_0$, using the continuity theorem once again. The hypothesis of the lemma is sufficient for the existence of a continuous inverse. (The proof of this analytical fact is similar to that of the familiar inverse function theorem for the $M = m$ case: The continuous derivatives make the transformation locally linear, and the rank condition suffices when the transformation is linear.)

The actual large sample form of $\bar{\theta}_n$ is given by the following theorem. The result is Fisher's, but a rigorous proof first appeared in Cramer's classic book [3] in 1946. The proof provides as a bonus an expression for $V_n(\theta_n)$ for general estimators θ_n. Denote by $B(\theta)$ the $M \times m$ matrix with (i,k)th entry

$$p_i(\theta)^{-\frac{1}{2}} \frac{\partial p_i(\theta)}{\partial \theta_k}$$

In analogy with the common $o(1)$ notation from analysis, $o_p(1)$ denotes any quantity converging in probability to zero as n increases. From this point, we shall for brevity omit the argument θ when $\theta = \theta_0$. Thus, for example, $B = B(\theta_0)$ and $V_n = V_n(\theta_0)$ in the statement of the following theorem.

THEOREM 3. *Under the conditions of Lemma 1, when* $G(\cdot|\theta_0)$ *holds,*

$$n^{\frac{1}{2}}(\bar{\theta}_n - \theta_0) = (B'B)^{-1}B'V_n + o_p(1). \tag{9}$$

PROOF. Since $\bar{\theta}_n$ is consistent and is assumed to attain the minimum of Q_n over Ω, and since $\partial Q_n/\partial\theta_k$ exists near θ_0, it follows that for sufficiently large n

$$\frac{\partial Q_n}{\partial \theta_k}(\bar{\theta}_n) = -\sum_{i=1}^{M} 2n \frac{N_i - np_i(\bar{\theta}_n)}{N_i} \frac{\partial p_i}{\partial \theta_k}(\bar{\theta}_n) = 0 \qquad k = 1,\ldots,m$$

with probability as near 1 as may be desired. Equivalently,

$$\sum_{i=1}^{M} \frac{N_i - np_i(\bar{\theta}_n)}{N_i^{\frac{1}{2}}} \left(\frac{n}{N_i}\right)^{\frac{1}{2}} \frac{\partial p_i}{\partial \theta_k}(\bar{\theta}_n) = 0 \qquad k = 1,\ldots,m. \tag{10}$$

We will apply the mean value theorem and the continuity theorem separately to the two factors in each summand of (10), remembering that $\bar{\theta}_n \to \theta_0(P)$ by Lemma 1. First,

$$\begin{aligned} N_i - np_i(\bar{\theta}_n) &= N_i - np_i - n(p_i(\bar{\theta}_n) - p_i) \\ &= N_i - np_i - n \sum_{k=1}^{m} \left[\frac{\partial p_i}{\partial \theta_k} + o_p(1)\right](\bar{\theta}_{nk} - \theta_{0k}). \end{aligned}$$

Combining this with

$$(np_i/N_i)^{\frac{1}{2}} = 1 + o_p(1) \tag{11}$$

establishes that $[N_i - np_i(\bar{\theta}_n)]/N_i^{\frac{1}{2}}$ is the ith component of the M-vector

$$V_n - Bn^{\frac{1}{2}}(\bar{\theta}_n - \theta_0) + o_p(1)n^{\frac{1}{2}}(\bar{\theta}_n - \theta_0) + o_p(1). \tag{12}$$

The second factor in (10) is similarly found to be

$$\left(\frac{n}{N_i}\right)^{\frac{1}{2}} \frac{\partial p_i}{\partial \theta_k}(\bar{\theta}_n) = p_i^{-\frac{1}{2}} \frac{\partial p_i}{\partial \theta_k} + o_p(1).$$

Equation (10) therefore becomes, in vector form,

$$B'V_n - [B'B + o_p(1)]n^{\frac{1}{2}}(\bar{\theta}_n - \theta_0) = o_p(1) \tag{13}$$

Now B'B is nonsingular by the rank m assumption, and since the determinant of a matrix is a continuous function of its elements,

$$\det(B'B + o_p(1)) \to \det(B'B) \neq 0(P).$$

Hence if A_n is the event that $B'B + o_p(1)$ is nonsingular, and χ_n the indicator function of this event, then the probability of A_n under $G(\cdot|\theta_0)$ approaches 1, $\chi_n \to 1(P)$, and $[B'B + o_p(1)]^{-1}\chi_n \to (B'B)^{-1}(P)$. Applying $[B'B + o_p(1)]^{-1}\chi_n$ to both sides of (13) gives

$$[B'B + o_p(1)]^{-1}\chi_n B'V_n - n^{\frac{1}{2}}(\bar{\theta}_n - \theta_0)\chi_n = o_p(1)$$

which in turn implies the result of the theorem.

Theorem 3 yields immediate fruit, and (12) will produce a later harvest as well. Reviewing the proof, it is easy to see that (12) holds for *any* consistent estimator θ_n of θ in the form

$$V_n(\theta_n) = V_n - Bn^{\frac{1}{2}}(\bar{\theta}_n - \theta_0) + o_p(1)n^{\frac{1}{2}}(\theta_n - \theta_0) + o_p(1). \tag{14}$$

This is the central relation in the theory of chi-square tests, as it expresses $V_n(\theta_n)$ in terms of the standardized multinomial vector V_n and a separate term reflecting the effect of estimating θ. Notice that the third term on the right is $o_p(1)$ whenever $n^{\frac{1}{2}}(\theta_n - \theta_0)$ converges in distribution. We can now provide quick proofs of several important results.

The first of these is Fisher's solution to the question of the behavior of Pearson's statistic when θ is estimated by $\bar{\theta}_n$. Substituting (9) into (14), we see that

$$V_n(\bar{\theta}_n) = (I_M - B(B'B)^{-1}B')V_n + o_p(1).$$

Asymptotic normality for V_n was immediate (see (7)). By the continuity theorem and the result from Section 2 on linear transformations of multivariate normal variables, it follows that under $G(\cdot|\theta_0)$,

$$V_n(\bar{\theta}_n) \to N_M(0,\Sigma)$$

$$\Sigma = (I_M - B(B'B)^{-1}B')(I_M - qq')(I_M - B(B'B)^{-1}B')$$

$$= I_M - qq' - B(B'B)^{-1}B'$$

The last equality is a consequence of the important relation $q'B = 0$, which holds because $\Sigma_{i=1}^M p_i = 1$ implies that $\Sigma_{i=1}^M \partial p_i/\partial\theta_k = 0$ for each k. The limiting null distribution of any statistic $V_n(\bar{\theta}_n)'C\,V_n(\bar{\theta}_n)$ is now given by Theorem 1. In particular, the Pearson statistic $P_n(\bar{\theta}_n) = V_n(\bar{\theta}_n)'V_n(\bar{\theta}_n)$ has the distribution of $\Sigma_{i=1}^M \lambda_i Z_i^2$ where λ_i are the characteristic roots of Σ. A bit of matrix multiplication will show that qq' and $B(B'B)^{-1}B'$ are symmetric idempotent matrices, that is, orthogonal projections. Moreover, we just saw that they are orthogonal to each other. Because qq' has rank 1 and $B(B'B)^{-1}B'$ has rank m by assumption, Σ is an orthogonal projection of rank M-m-1. So its characteristic roots are M-m-1 1's and m+1 0's, and the limiting null distribution of $P_n(\bar{\theta}_n)$ is $\chi^2(M\text{-}m\text{-}1)$. Notice especially that this is true for any θ_0 in Ω, even though Σ varies with θ_0. This is the famous "subtract one degree of freedom for each parameter estimated" result.

Now $\bar{\theta}_n$ is often not the most convenient available estimator of θ. In testing fit to the univariate normal family with $\theta = (\mu,\sigma)$, for example, the cell probability for a cell $E_i = (a_{i-1},a_i]$ is

$$p_i(\mu,\sigma) = \Phi\left(\frac{a_i-\mu}{\sigma}\right) - \Phi\left(\frac{a_{i-1}-\mu}{\sigma}\right),$$

where Φ is the standard normal df. The equations (10) have no closed-form solution, nor do the yet more complicated equations $\partial P_n(\theta)/\partial\theta_k = 0$ defining

the minimum chi-square estimator. A visit to your local computing center will uncover library programs for evaluating Φ and iteratively solving the equations (10). Nonetheless, it is hard to ignore the universally used sample mean and variance, $\hat{\theta}_n = (\bar{X}, s)$. What will befall us if we use these estimators in the Pearson statistic instead of $\bar{\theta}_n$? To answer this question, we must discover the large-sample behavior of $\hat{\theta}_n$ and then consult (14).

This is best done in greater generality. The sample mean and standard deviation (taking $s^2 = \Sigma_{j=1}^{n}(X_j - \bar{X})^2/n$) form the *maximum likelihood estimator* (MLE) of $\theta = (\mu, \sigma)$ in the univariate normal family. In general, the MLE $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ of θ is defined as any value of θ maximizing the joint density function of the observations considered as a function of θ for given $X_1, \ldots, X_n$. This recipe for a general method of estimating parameters is another of Fisher's contributions. It is intuitively forceful, estimating θ to be the value making the actually observed $X_1, \ldots, X_n$ "most probable." More satisfying to the perverse theoretician, the MLE is guaranteed to have good properties in large samples. Specifically, suppose that the X_j are independent with common density function $g(\cdot|\theta_0)$. Then under reasonable smoothness conditions,

$$(15) \qquad n^{\frac{1}{2}}(\hat{\theta}_n - \theta_0) = J(\theta_0)^{-1} n^{-\frac{1}{2}} \sum_{j=1}^{n} \frac{\partial \log f(X_i|\theta_0)}{\partial \theta} + o_p(1).$$

Here $\partial \log f/\partial\theta$ is the m-vector of partial derivatives with respect to $\theta_1, \ldots, \theta_m$ and $J(\theta)$ is the m×m matrix with (k,ℓ)th component

$$E_\theta\left[\frac{\partial \log f(X|\theta)}{\partial \theta_k} \; \frac{\partial \log f(X|\theta)}{\partial \theta_\ell}\right].$$

It follows from (15) by the multivariate central limit theorem that

$$(16) \qquad n^{\frac{1}{2}}(\hat{\theta}_n - \theta_0) \to N_m(0, J(\theta_0)^{-1}).$$

The matrix $J(\theta)$ is called the *information matrix* for the family $f(x|\theta)$. The inverse $J(\theta_0)^{-1}$ is the "smallest possible" covariance matrix for the limiting distribution of an estimator of θ, in several specific senses which this is not the place to specify. Thus (16) says roughly that the MLE has the tightest possible concentration about the true θ_0 in large samples. This is called *asymptotic efficiency* of the MLE.

In the light of this pleasing result, it would be very intelligent, if we wish to estimate θ from cell frequencies, to apply the MLE recipe to the indicator variables $\delta_1,\ldots,\delta_n$ indicating into which cells $X_1,\ldots,X_n$ fall. A bit of work shows that the information matrix in this case is B'B, and that (15) reduces to (9). *The minimum chi-square and minimum modified chi-square and maximum likelihood estimators are all asymptotically equivalent ways of estimating θ from the cell frequencies.* That's aesthetically satisfying.

Having summed up half a century of hard work on MLE's in one paragraph, we can now substitute (15) into (14). Here $\hat{\theta}_n$ is the MLE of θ from the ungrouped observations $X_1,\ldots,X_n$, not the less efficient MLE based on the cell frequencies. Fortune is with us. The first term in (14), namely V_n, was expressed at the beginning of this section as a sum of n terms, one for each X_i. The second term, namely (15), has the same form. And the rest of (14) is $o_p(1)$. So we obtain from the first two terms a sum which is asymptotically normal by the multivariate central limit theorem. A computation of covariances gives specifically that

$$V_n(\hat{\theta}_n) \xrightarrow{\mathcal{D}} N_M(0,\Sigma)$$

$$\Sigma = I_M - qq' - BJ^{-1}B'.$$

Therefore the limiting null distribution of $P_n(\hat{\theta}_n)$ is that of $\Sigma_{i=1}^{M}\lambda_i Z_i^2$, where λ_i are the characteristic roots of Σ.

Now $BJ^{-1}B'$ has the same rank m as does B, and therefore has as its range the range of B. Since qq' is an orthogonal projection of rank 1 orthogonal to B, the characteristic roots of Σ include M-m-1 1's (Σ acts as the identity in directions orthogonal to the direct sum of the ranges of B and qq') and one 0 (Σ acts as zero on the range of qq'). The remaining roots $\lambda_1,\ldots,\lambda_m$ reflect the fact that Σ acts as $I_M - BJ^{-1}B'$ on the range of B. One version of the "efficiency" of the MLE is that $\hat{\theta}_n$ is asymptotically preferable to $\bar{\theta}_n$ in the sense that J-B'B is nonnegative definite. From this it can be shown by matrix mangling that $0 \leq \lambda_i < 1$, and $0 < \lambda_i < 1$ except in the unusual case when J-B'B fails to be positive definite. The λ_i of course depend on θ_0, as well as on the specific hypothesized family $g(\cdot|\theta)$.

We have now reached the second major consequence of (14). The statistic $P_n(\hat{\theta}_n)$ has as its limiting null distribution the distribution of

$$\chi^2(M\text{-}m\text{-}1) + \sum_{i=1}^{m} \lambda_i Z_i^2. \tag{17}$$

This is *not* a chi-square distribution. What is worse, the distribution varies with θ_0, so there is no single limiting distribution across the composite null hypothesis. Since $0 \leq \lambda_i < 1$ for all θ_0, it is at least true that critical points of (17) lie between those of $\chi^2(M\text{-}m\text{-}1)$ and $\chi^2(M\text{-}1)$. When there are many cells and few parameters, these bounds are close together. But we cannot without care follow such natural paths as the use of $\bar{X}$ and s in the Pearson statistic to test for normality.

After Chernoff and Lehmann [2] obtained the result (17) in 1954, statistical theory produced a variety of ways of escape. One is suggested immediately by Theorem 2: Compute a generalized inverse of Σ and use the

corresponding quadratic form. It is easy to see from our previous study of Σ that Σ has rank M-1 and

$$\Sigma^{-} = (I_M - BJ^{-1}B')^{-1}$$

whenever J - B'B is positive definite. If now

$$C_n = (I_M - B(\hat{\theta}_n)J(\hat{\theta}_n)^{-1}B(\hat{\theta}_n))^{-1}$$

then $C_n \to \Sigma^{-}(P)$ and

$$V_n(\hat{\theta}_n)'C_nV_n(\hat{\theta}_n) \to \chi^2(M-1)$$

under $G(\cdot|\theta)$ for any θ in Ω. This statistic is not as hard to compute as may appear, as will be shown by example in Section 7. This statistic was first studied by Rao and Robson [5], but without the supporting theory. Rao and Robson present C_n in the form

$$C_n = I_M + B(\hat{\theta}_n)[J(\hat{\theta}_n) - B(\hat{\theta}_n)'B(\hat{\theta}_n)]^{-1}B(\hat{\theta}_n)'$$

which makes it clear that the new statistic $V_n(\hat{\theta}_n)'C_nV_n(\hat{\theta}_n)$ is the Pearson statistic plus a second quadratic form. Challenge: prove that the two expressions given for C_n are equivalent.

If $P_n(\hat{\theta}_n)$ can be built up to reach $\chi^2(M-1)$, it can also be chopped down to $\chi^2(M-m-1)$. Since $B(B'B)^{-1}B'$ is the orthogonal projection onto the range of B, you should be able to show that $V_n(\hat{\theta}_n)'D(\hat{\theta}_n)V_n(\hat{\theta}_n)$ has the $\chi^2(M-m-1)$ limiting null distribution, where

$$D(\theta) = I_M - B(\theta)[B'(\theta)B(\theta)]^{-1}B'(\theta).$$

This result does not even depend on the use of $\hat{\theta}_n$; $\bar{\theta}_n$ and most other estimators of θ give the same result. But the price of such generality is inefficiency. Simulations suggest that the $D(\hat{\theta}_n)$ statistic often has low

power (that is, little ability to detect that the null hypothesis is false). The C_n statistic, on the other hand, is usually more powerful than the Pearson statistic. It deserves consideration as a standard chi-square test for goodness of fit.

5. CONTINGENCY TABLES

The use of chi-square statistics for testing fit is based on creating a set of multinomial observations by counting cell frequencies. Because only cell frequencies are used in the tests, some information is lost. There are other classes of tests of fit which are generally more powerful than chi-square tests, though none so flexible and widely applicable. There are, however, situations in which multinomial observations arise naturally. In such cases, chi-square tests are the natural large sample tests. A common instance is a *contingency table:* sample units are categorized according to two or more variables with the intent of discovering the relationship between the variables. The data consist of the frequencies of sample units in all possible cross-classifications. Here is the layout of a 2×s contingency table, with the notation used for the cell frequencies.

(18)

N_{11}	N_{12}	...	N_{1s}	$N_{1\cdot}$
N_{21}	N_{22}	...	N_{2s}	$N_{2\cdot}$
$N_{\cdot 1}$	$N_{\cdot 2}$		$N_{\cdot s}$	

We have used the common notation in which a dot replaces an index when the frequencies are summed over the full range of that index. Thus $N_{\cdot j}$ is the jth column sum, the total number of units which fell in category j for the column variable. For simplicity, this 2×s table will be the focus of our discussion, though the conclusions are generally valid.

What is the proper probability model for these data? *The model must reflect the way in which the data were collected.* There are several different sampling procedures which could lead to the table (18). A single random sample of size n might be selected, then categorized in two ways. For example, a random sample of persons being treated for cancer might be classified by sex (2 categories) and type of cancer (s categories). Call this Model A. Under Model A, the cell frequencies N_{ij} have a single 2s-nomial distribution. The marginal frequencies are all random, and satisfy

$$N_{1.} + N_{2.} = \sum_{j=1}^{s} N_{.j} = \sum_{i=1}^{2} \sum_{j=1}^{s} N_{ij} = n. \tag{19}$$

Table (18) might also result from selecting two independent random samples, of male cancer patients and female cancer patients separately, then categorizing each patient by type of cancer. Under this Model B, table (18) contains two independent s-nomial distributions. Although (19) still holds, $N_{1.}$ and $N_{2.}$ are no longer random, for they are the sample sizes chosen by the experimenter. Model C reverses the roles of the variables: choose independent random samples of patients under treatment for each of s types of cancer, then categorize each by sex. Here there are s independent binomials, and the $N_{.j}$ are nonrandom sample sizes.

All three models for table (18) are sets of independent multinomial observations. Chi-square methods provide tests of hypotheses concerning the cell probabilities in any such setting. This is a generalization of the situation arising in tests of fit, where only a single multinomial sample was available, but the theory of chi-square tests follows much the same line.

Hypotheses for these models are stated in terms of the cell probabilities p_{ij} for the sampled population. Each model imposes different constraints on the p_{ij}. Model A requires only that

$$\sum_{i=1}^{2} \sum_{j=1}^{s} p_{ij} = 1 \tag{20}$$

(and of course that $0 \leq p_{ij} \leq 1$ for all i and j). Model B states that

$$\begin{aligned} p_{1\cdot} &= \sum_{j=1}^{s} p_{1j} = 1 \\ p_{2\cdot} &= \sum_{j=1}^{s} p_{2j} = 1, \end{aligned} \tag{21}$$

since in this case two independent s-nomials are observed. Model C assumes instead that $p_{\cdot j} = p_{1j} + p_{2j} = 1$ for each j. The most common hypotheses (and the only ones we will consider) formalize the statement that there is no connection between the two categorizations - in the example, no connection between the sex of a cancer patient and the type of cancer under treatment. In Model A, this is the hypothesis of *independence*,

$$H_A: \ p_{ij} = p_{i\cdot}p_{\cdot j} \qquad i = 1, 2 \text{ and } j = 1,\ldots,s. \tag{22}$$

In Model B, the hypothesis is that of *two identical s-nomial distributions*,

$$H_B: \ p_{1j} = p_{2j} \qquad j = 1,\ldots,s. \tag{23}$$

For Model C, no connection between categorizations is expressed as the hypothesis of *s identical binomial distributions*,

$$H_C: \ p_{11} = p_{12} = \ldots = p_{1s}.$$

In all of this, our concern has been simply to translate the sampling design and the question to be asked of the data into a mathematical model.

This process is often less clear and more controversial than the theory which follows from the model selected. There are, for example, yet other models for the data of table (18). These models assume that the data arise not as random samples from a large population, but from experimental randomization of a finite set of experimental units. In the randomization analog of Model B, n units (say lab rats) are available, of which $N_{1\cdot}$ are assigned to Treatment 1 and $N_{2\cdot}$ to Treatment 2 by random allocation. The response of each rat falls into one of s categories, and N_{ij} is the number of rats receiving Treatment i with response j resulting. Just as in Model B, (19) holds, $N_{1\cdot}$ and $N_{2\cdot}$ are fixed, and the hypothesis to be tested is that of equal response distributions. But because the responses of the fixed set of rats are dependent, the distribution of the N_{ij} is no longer multinomial. Moreover, under the null hypothesis of "no treatment effect", the number $N_{\cdot j}$ of rats showing response j is a nonrandom characteristic of the particular set of rats used in the experiment.

Now it turns out that under Model B the *conditional* null distribution of the N_{ij} given the observed values of $N_{\cdot j}$ ($j = 1,\ldots,s$) is exactly the null distribution of the N_{ij} under the randomization model just described. Because such experimental randomization is common practice in many fields of work, it has been the historical pattern to argue that the randomization models are "exact" and the multinomial models (and therefore the chi-square tests) are valid only when interpreted conditionally. But wait - since the randomization model considers only the fixed set of units actually involved in the experiment, any inference based on that model can apply only to those particular units. If our n rats are in some way atypical of rats-in-general, this will influence the outcome of the experiment, and no conclusions can be drawn for rats-in-general. In practice, we argue or assume that our particular units are representative of some larger population. That is, we

commonly act as if we had samples from a population of interest. What is more, steps are often taken to justify this assumption - we cannot sample the population of all rats or all cancer patients, but we can select our experimental units at random from a large pool of accessible units. In such a case neither the randomization nor the multinomial models are strictly appropriate, but the multinomial models do represent the conditions which the selection and allocation of units aim to attain.

This discussion is not at all a digression. It is rather a paradigm of the features which distinguish areas of applied mathematics (such as statistics) from mathematics-for-its-own-sake. It is time, however, to assume that one of the multinomial models adequately describes the data and to turn to tests of H_A, H_B or H_C. If we wish to detect any deviation from the null hypothesis (not just deviations in some specified direction), an omnibus test is in order. And if the sample sizes are moderately large, chi-square methods provide such tests.

We will follow the pattern of Section 4, denoting the unknown vector of parameters by θ. In the contingency table case, θ consists of a set of cell probabilities which determine all of the p_{ij} when combined with the constraints imposed by the model and by the null hypothesis. Under Model B, for example, we will take $\theta = (p_{11},\ldots,p_{1,s-1})'$, since (21) and (23) then determine the complete set of cell probabilities. The estimators, test statistics, and limiting distributions discussed below do not depend on this particular choice of s-1 cell probabilities for θ, but the dimension $m = s-1$ of θ is a consequence of the model and the hypothesis.

The probability function in Model B is

$$\frac{N_1\,!}{N_{11}!\ldots N_{1s}!}\prod_{j=1}^{s} p_{1j}^{N_{1j}}\,\frac{N_2\,!}{N_{21}!\ldots N_{2s}!}\prod_{j=1}^{s} p_{2j}^{N_{2j}}\,.$$

To estimate θ by the maximum likelihood method, express each p_{ij} as $p_{ij}(\theta)$ and the probability function for given N_{ij} as a function $L(\theta)$ of θ. Then solving

$$\frac{\partial \log L(\theta)}{\partial \theta_k} = \frac{N_{1k}}{p_{1k}} - \frac{N_{1s}}{p_{1s}} + \frac{N_{2k}}{p_{1k}} - \frac{N_{2s}}{p_{1s}} = 0 \qquad k = 1,\ldots,s-1$$

(recall $\theta_k = p_{1k}$) produces the MLE

$$\hat{p}_{1k} = \frac{N_{1k} + N_{2k}}{n} = \frac{N_{\cdot k}}{n} \qquad k = 1,\ldots,s-1.$$

This is the "obvious" estimator of p_{1k} under H_B, namely the overall relative frequency of the kth response in the two samples. Another way to describe $\hat{p}_{1k}$ is as the weighted arithmetic mean of the relative frequencies $N_{1k}/N_{1\cdot}$ and $N_{2k}/N_{2\cdot}$ of the kth response in the separate samples. The Pearson chi-square statistic for two independent s-nomials is

$$\sum_{j=1}^{s} \frac{[N_{1j} - np_{1j}(\hat{\theta})]^2}{np_{1j}(\hat{\theta})} + \sum_{j=1}^{s} \frac{[N_{2j} - np_{2j}(\hat{\theta})]^2}{np_{2j}(\hat{\theta})}$$

$$= \sum_{i=1}^{2}\sum_{j=1}^{s} \frac{[N_{ij} - N_{i\cdot}N_{\cdot j}/n]^2}{N_{i\cdot}N_{\cdot j}/n}\,.$$

When the p_{ij} are known, this statistic has (s-1) + (s-1) = 2s - 2 degrees of freedom. Here, m = s-1 parameters were estimated by the multinomial MLE method, so since (2s-2) - (s-1) = s-1, the limiting null distribution is $\chi^2(s-1)$.

If the data of table (18) arose from a single random sample (Model A), the probability function is

$$\prod_{i=1}^{2} \prod_{j=1}^{s} \frac{n!}{N_{ij}!} p_{ij}^{N_{ij}} .$$

The unknown parameter can be taken to be $\theta = (p_{11},\ldots,p_{1s})'$ since (20) and (22) then determine the full set of cell probabilities. Once again other choices of θ are possible but all have $m = s$. Computing the MLE of θ gives the natural estimator for Model A under H_A, namely

$$\hat{p}_{1k} = \frac{N_{1\cdot}}{n} \frac{N_{\cdot k}}{n} .$$

(Compare (22) to see why this is the natural estimator.) The Pearson chi-square statistic for this single 2s-nomial model is

$$\sum_{i=1}^{2} \sum_{j=1}^{s} \frac{[N_{ij} - np_{ij}(\hat{\theta})]^2}{np_{ij}(\hat{\theta})} = \sum_{i=1}^{2} \sum_{j=1}^{s} \frac{[N_{ij} - N_{i\cdot}N_{\cdot j}/n]^2}{N_{i\cdot}N_{\cdot j}/n}$$

and has $2s-m-1 = s-1$ degrees of freedom.

Look closely. *The Pearson chi-square statistics for testing* H_A *in* Model A *and for testing* H_B *in* Model B *are identical, and have the same* $\chi^2(s-1)$ *limiting null distribution.* And of course the same statistic results from testing H_C in Model C. This serendipitous outcome depends very much on the fact that maximum likelihood estimation was used. As Section 4 proclaimed, asymptotically equivalent statistics can be obtained by using either the minimum chi-square or the minimum modified chi-square method to estimate θ. But only asymptotically equivalent. Let us apply the minimum modified chi-square method to Model B. We must choose $\theta = (p_{11},\ldots,p_{1,s-1})'$ to minimize the modified chi-square statistic

$$\sum_{j=1}^{s} \frac{[N_{1j} - N_{1\cdot}p_{1j}(\theta)]^2}{N_{1j}} + \sum_{j=1}^{s} \frac{[N_{2j} - N_{2\cdot}p_{2j}(\theta)]^2}{N_{2j}} .$$

Differentiation followed by a short, ugly calculation shows that the minimum modified chi-square estimator of p_{1k} is proportional to the weighted

harmonic mean of the relative frequencies $N_{1k}/N_{1\cdot}$ and $N_{2k}/N_{2\cdot}$ for the kth response. While not entirely outrageous, this is surely less appealing than the MLE result. In Model A, the situation is worse: the equations arising from differentiating the modified chi-square statistic are nonlinear, and have no closed-form solution. The Pearson statistics for Models A and B when minimum modified chi-square estimators are used are *not* identical. The minimum chi-square estimators have no explicit expressions in either model, and again the chi-square statistics differ. No wonder the MLE is always used for contingency tables.

There is a pattern to the use of these latter estimation methods in hypothesis testing for independent multinomial observations. The minimum modified chi-square method produces a set of *linear* equations to be solved for the estimated parameters whenever the hypothesis is linear in the cell probabilities. This was true of H_B but not of H_A. In many situations it is easier to compute minimum modified chi-square estimators than the MLE's. Minimum chi-square estimators, on the other hand, can rarely be obtained in closed form and are seldom used.

6. A FURTHER RANGE

This survey of chi-square tests has entirely ignored several areas of considerable interest to users of these tests. Computers make it feasible to obtain the exact distributions of the test statistics in small samples, both for use and for assessment of the accuracy of the chi-square approximations. The relative power of the tests can be studied either by calculation and simulation or, in large samples, by various mathematical devices. I have chosen to restrict this essay to the study of large sample distribution theory under the null hypothesis. Even here there is a further range of

theory which both opens up new possibilities for the user and illustrates the use of increasingly sophisticated mathematics in statistical theory.

We have been assuming almost without reflection that the number of cells M in a chi-square test of fit remains fixed as the sample size increases, and that the cells E_i are fixed without regard to the data. Neither assumption is necessarily realistic as a description of statistical practice. It is common to use more cells when blessed with a larger sample, and equally common (though less publicly admitted) to move the cells to the data. What are the consequences of incorporating these innovations in the chi-square statistics of Section 4?

Increasing the number of cells M as the sample size n increases has radical consequences. When M grows with n, the Pearson statistic for testing fit to a completely specified distribution has a *normal*, not a chi-square, limiting null distribution when properly standardized. This is in accord with intuition, since the $\chi^2(M-1)$ limiting distribution of the Pearson statistic approaches normality as $M \to \infty$. (Apply the central limit theorem to $\Sigma_1^{M-1} Z_i^2$.) One expects that the limiting null distribution when parameters are estimated will also be normal, with a different standardization perhaps required. No proof of this has been given. The lack of attention to this problem may be due in part to simulations suggesting that replacing M by $M(n) \to \infty$ produces tests which compare favorably with fixed-M chi-square tests only against very short-tailed alternatives.

The second innovation, use of data-dependent cells, has been better studied and is finding its way increasingly into practical use. Suppose then

that the cells E_i are replaced by

$$E_{in}(X_1,\ldots,X_n) \qquad i = 1,\ldots,M$$

in the general chi-square statistic (6) of Section 4. For simplicity we consider only univariate observations X_j and cells $E_{in} = (a_{i-1,n}, a_{in}]$ which are intervals with endpoints $a_{in} = a_{in}(X_1,\ldots,X_n)$. It is only reasonable to demand that the random cells settle down as the sample size increases,

$$a_{in} \to a_{i0} = a_{i0}(\theta_0) \ (P) \qquad \text{under } G(\cdot|\theta_0).$$

An example of useful data-dependent cell boundaries is $a_{in} = \bar{X}_n + c_i s_n$ in testing fit to the univariate normal family. The sample mean $\bar{X}_n$ moves the cells to the data, and the sample standard deviation s_n adjusts the cell widths to the dispersion of the data. Here $a_{i0}(\mu,\sigma) = \mu + c_i\sigma$ are the limiting cell boundaries.

If a_n denotes the vector of cell boundaries ($a_{0n} \equiv -\infty$, $a_{Mn} \equiv +\infty$), then the "cell probabilities" under the null hypothesis are now

$$p_i(\theta,a_n) = \int_{a_{i-1,n}}^{a_{in}} dG(x|\theta) = G(a_{in}|\theta) - G(a_{i-1,n}|\theta).$$

The M-vector of standardized cell frequencies becomes $V_n(\theta,a_n)$ with ith component

$$\frac{N_i(a_n) - np_i(\theta,a_n)}{[np_i(\theta,a_n)]^{\frac{1}{2}}}$$

where $N_i(a_n)$ is the number of $X_1,\ldots,X_n$ in E_{in}. But *the cell frequencies* $N_i(a_n)$ *are no longer multinomial*, since the cell boundaries are dependent on the observations X_j being counted. The central mathematical hurdle of establishing asymptotic normality of $V_n(\theta_n,a_n)$ for estimators θ_n and random cell boundaries a_n is now much more difficult. Fortunately, there is an

elegant modern technique leading to a pleasing result.

The pleasing result is that the asymptotic distribution of $V_n(\theta_n, a_n)$ under $G(\cdot|\theta_0)$ is the same as that of $V_n(\theta_n, a_0)$. That is, *the asymptotic behavior of any random-cell chi-square statistic is exactly that of the same statistic using the limiting cell boundaries* a_{i0}. Speaking roughly, the use of data-dependent cells has no effect. The naive user who moves his cells to the data is safe. Actually, the dependence of the limiting cells on the (unknown) true θ_0 complicates these rough conclusions slightly. But any of the statistics in Section 4 which has a θ_0-free limiting null distribution using fixed cells has that same distribution using any set of converging random cells.

What of the promised elegant modern technique? Since $V_n(\theta_n, a_0)$ is just our old acquaintance $V_n(\theta_n)$ for a particular set of fixed cells, let us use the latter notation. We must show that under $G(\cdot|\theta_0)$

$$V_n(\theta_n, a_n) - V_n(\theta_n) = o_p(1).$$

Since both $p_i(\theta_n, a_n)$ and $p_i(\theta_n)$ converge in probability to $p_i(\theta_0)$ whenever θ_n is a consistent sequence of estimators and p_i is continuous, their presence in the denominators can be ignored. We need only prove that the expression

$$n^{-\frac{1}{2}}[N_i(a_n) - np_i(\theta_n, a_n)] - n^{-\frac{1}{2}}[N_i - np_i(\theta_n)] \tag{24}$$

is $o_p(1)$.

Introduce the *empiric distribution function*

$$G_n(t) = n^{-1}\{\text{Number of } X_1, \ldots, X_n: \ X_j \leq t\}.$$

This is the natural estimator of the common df of the X_j. It increases from 0 to 1 in jumps of 1/n at each observation. At any fixed t, $G_n(t)$ is

a multiple of a binomial random variable with success probability $G(t|\theta_0)$ when this is the true df of the X_j. So the *empiric distribution function process*

$$W_n(t) = n^{\frac{1}{2}}\{G_n(t) - G(t|\theta_0)\}$$

has a normal limiting distribution for each fixed t. Since

$$n^{-\frac{1}{2}}N_i(a_n) = n^{\frac{1}{2}}\{G_n(a_{in}) - G_n(a_{i-1,n})\}$$

$$p_i(\theta,a_n) = G(a_{in}|\theta) - G(a_{i-1,n}|\theta),$$

there is some hope of expressing (24) in terms of W_n and using the convergence properties of that process to achieve our goal.

In Section 3 we saw that convergence in distribution for random variables amounted to describing a random variable Y_n by a probability measure F_n on R^p and defining convergence as convergence of the measures $F_n(A)$ of all Borel sets A having boundaries with measure 0 under the limiting distribution. This development extends at once to more general spaces than R^p. Indeed, such an extension is the primary topic of Billingsley's book [1] which was cited in Section 3. Now a stochastic process such as W_n is a *random function* - a function of the real variable t which varies with the underlying probability mechanism generating the X_j. Just as a random variable can be identified with a probability measure on R^p, so a random function can be identified with a probability measure on a suitable function space. Convergence in distribution for processes then has the same definition as convergence in distribution for random variables. This viewpoint, adopted from functional analysis, has become a standard tool of statistical large sample theory. Billingsley's book is a basic exposition, restricted to metric spaces, Borel σ-fields, and

random functions of a single real variable.

It turns out that $W_n(t)$ does converge in distribution to a process $W_0(t)$ which is a variation of Brownian motion, one of the most familiar stochastic processes. This is an analog of the central limit theorem. Of this powerful result we need only two details: $W_0(t)$ is continuous with probability 1, and the function space on which W_n and W_0 are probability measures has the property that convergence to a continuous limit function is always uniform.

The machinery to crush (24) is now assembled. Arithmetic shows that (24) is

$$\begin{aligned}(25)\qquad &\{W_n(a_{in}) - W_n(a_{i0})\} - \{W_n(a_{i-1,n}) - W_n(a_{i-1,0})\} \\ &+ n^{\frac{1}{2}}\{p_i(a_n,\theta_0) - p_i(a_n,\theta_n)\} - n^{\frac{1}{2}}\{p_i(a_0,\theta_0) - p_i(a_0,\theta_n)\}\end{aligned}$$

Applying the mean value theorem to the last two terms in (25) gives

$$[\frac{\partial p_i}{\partial \theta}(\theta_n^*, a_n) - \frac{\partial p_i}{\partial \theta}(\theta_n^{**}, a_0)]' n^{\frac{1}{2}}(\theta_n - \theta_0)$$

for θ_n^*, θ_n^{**} between θ_n and θ. This is $o_p(1)$ whenever the m-vector of derivatives $\partial p_i/\partial\theta$ is continuous and $n^{\frac{1}{2}}(\theta_n - \theta_0)$ converges in distribution. The first two terms in (25) have the form

$$W_n(c_n) - W_n(c)$$

where $c_n \to c(P)$. Now the two general facts of Section 3 apply to convergence in distribution of processes as well. So

$$(W_n, c_n) \xrightarrow{\mathcal{D}} (W_0, c)$$

by the second general fact. The function

$$\varphi(f,t) = f(t) - f(c)$$

is continuous with probability 1 with respect to the distribution of (W_0,c). (Check that: if $f_n \to f$ *uniformly* and $t_n \to c$, then $\varphi(f_n,t_n) \to \varphi(f,c) = 0$. That implies continuity with probability 1 because $W_0(t)$ is continuous, and convergence to a continuous limit function is uniform in the function space at hand.) So by the continuity theorem

$$W_n(c_n) - W_n(c) = \varphi(W_n,c_n) \xrightarrow{\mathcal{D}} \varphi(W_0,c) \equiv 0.$$

Convergence in distribution to a constant is convergence in probability, so we have shown that (25) is $o_p(1)$.

This essentially simple argument can be generalized to multivariate observations and alternative hypotheses. A full treatment appears in [4]. The examples in the next section illustrate the advantages of being free to use data-dependent cells.

7. SOME EXAMPLES

In this section the results of Sections 4 and 6 will be applied to produce several chi-square tests of fit to the family of exponential densities

$$g(x|\theta) = \frac{1}{\theta} e^{-x/\theta} \qquad 0 < x < \infty$$

(26)

$$\Omega = \{\theta: \ 0 < \theta < \infty\}.$$

There are many tests of fit for so standard a family which exceed the chi-square tests in power. Chi-square tests of fit have their greatest potential usefulness in situations where other tests of fit cannot be used (discrete or multivariate data), or where the work of computing critical points for a non-tabled distribution is not justified. But restricting

ourselves to a single, simple hypothesized family has obvious expository advantages.

Suppose then that $X_1,\ldots,X_n$ are independent random variables having a common unknown distribution which is hypothesized to belong to the family (26).

The Pearson statistic. Choose M fixed cells $E_i = (a_{i-1},a_i]$ partitioning $(0,\infty)$. The cell probabilities under the null hypothesis are

$$p_i(\theta) = \int_{a_{i-1}}^{a_i} \frac{1}{\theta} e^{-x/\theta} dx = a_{i-1}e^{-a_{i-1}/\theta} - a_i e^{-a_i/\theta}.$$

We will estimate θ by the grouped data MLE $\bar{\theta}_n$. The equation resulting from differentiating the logarithm of the multinomial probability function of $N_1,\ldots,N_M$ with respect to θ is

$$\sum_{i=1}^{M} N_i \frac{a_{i-1}e^{-a_{i-1}/\theta} - a_i e^{-a_i/\theta}}{e^{-a_{i-1}/\theta} - e^{-a_i/\theta}} = 0. \tag{27}$$

This has no closed-form solution, but is easily solved iteratively to obtain $\bar{\theta}_n$. Substituting this numerical value into the Pearson statistic

$$\sum_{i=1}^{M} \frac{[N_i - np_i(\bar{\theta}_n)]^2}{np_i(\bar{\theta}_n)}$$

gives a test statistic with approximately the $\chi^2(M-2)$ null distribution.

Using the raw data MLE. The maximum likelihood estimator of θ from $X_1,\ldots,X_n$ is the sample mean, $\hat{\theta}_n = \bar{X}_n$. Who would wish to solve (27) when $\bar{X}_n$ will do our estimating? If $\bar{X}_n$ is used in the Pearson statistic, a θ-dependent limiting null distribution results. But a nice feature of random cells now appears: in testing fit to location-parameter and scale-parameter families, random cells can eliminate the θ-dependence of the null distribution. In this case, we use cells

$$E_i(\bar{X}_n) = (c_{i-1}\bar{X}_n,\ c_i\bar{X}_n]$$

where the c_i are constants. In the notation of Section 6, $a_{in} = c_i\bar{X}$ and

$$p_i(\theta,a_n) = \int_{c_{i-1}\bar{X}}^{c_i\bar{X}} \theta^{-1}e^{-x/\theta}dx = e^{-c_{i-1}\bar{X}/\theta} - e^{-c_i\bar{X}/\theta}. \tag{28}$$

The estimated cell probabilities $p_i(\bar{X},a_n) = e^{-c_{i-1}} - e^{-c_i}$ do not depend on the sample! The Pearson statistic for random cells of this form is algebraically unchanged by the transformation $X_j \to X_j/\theta$ because the cell boundaries move in such a way as to keep the cell frequencies as well as the estimated cell probabilities fixed. Since X_j/θ has the $g(\cdot|1)$ density function when X_j has the $g(\cdot|\theta)$ density function, the distribution of the statistic does not depend on θ. If in particular $c_i = -\log(1 - \frac{i}{M})$, we obtain M equiprobable cells, $p_i \equiv 1/M$.

The use of random cells thus produces a θ-free null distribution. Not only that, the choice of equiprobable cells simplifies the computation of the statistic and has been shown to have good power properties when fit to a single distribution is being tested. But the limiting distribution is *not* chi-square. It is the distribution of

$$\chi^2(M-2) + \lambda Z_1^2,$$

and requires a special computation to obtain critical points even though λ does not depend on θ.

Using a different quadratic form. Both of the statistics we have thus far applied to testing fit to the family (26) have disabilities in ease of use. The C_n statistic of Rao and Robson can be explicitly computed, has a $\chi^2(M-1)$ limiting distribution, and also appears from simulations to be more powerful than its two competitors. Let us find its form for this problem. The general forms of both the C_n and $D(\hat{\theta}_n)$ statistics from Section 4 simplify

because $\Sigma_1^M \partial p_i/\partial\theta_k = 0$ implies that

$$V_n'B = n^{-\frac{1}{2}}\left(\sum_{i=1}^{M} \frac{N_i}{p_i} \frac{\partial p_i}{\partial\theta_1}, \ldots, \sum_{i=1}^{M} \frac{N_i}{p_i} \frac{\partial p_i}{\partial\theta_m}\right).$$

When m = 1, the Rao-Robson statistic reduces to

$$R_n = \sum_{i=1}^{M} \frac{(N_i - np_i)^2}{np_i} + \frac{1}{nD}\left(\sum_{i=1}^{M} \frac{N_i}{p_i} \frac{dp_i}{d\theta}\right)^2$$

where

$$D = J - \sum_{i=1}^{M} \frac{1}{p_i}\left(\frac{dp_i}{d\theta}\right)^2$$

and J, p_i, $dp_i/d\theta$ are all evaluated at $\theta = \hat{\theta}_n$.

The use of equiprobable random cells continues to have advantages in simplicity and (probably) in power, so the cells $(c_{i-1}\bar{X}, c_i\bar{X}]$ for $c_i = -\log(1 - \frac{i}{M})$ will again be employed. From (28),

$$\left.\frac{dp_i}{d\theta}\right|_{\theta=\bar{X}} = \left.\frac{c_{i-1}\bar{X}}{\theta^2} e^{-c_{i-1}\bar{X}/\theta} - \frac{c_i\bar{X}}{\theta^2} e^{-c_i\bar{X}/\theta}\right|_{\theta=\bar{X}}$$

$$= \frac{1}{\bar{X}}\left(c_{i-1}e^{-c_{i-1}} - c_i e^{-c_i}\right)$$

and a short calculation shows that the test statistic is

$$\frac{M}{n}\sum_{i=1}^{M}\left(N_i - \frac{n}{M}\right)^2 + \frac{M^2}{n}\frac{1}{(1 - M\Sigma_1^M d_i^2)}\left\{\sum_{i=1}^{M}\left(N_i - \frac{n}{M}\right)d_i\right\}^2.$$

Here $d_i = c_{i-1}e^{-c_{i-1}} - c_i e^{-c_i}$ and N_i is the number of $X_1, \ldots, X_n$ in $(c_{i-1}\bar{X}, c_i\bar{X}]$. This is the recommended chi-square statistic for this problem.

Censored data. It is quite common in experiments on reliability or survival time not to wait for all the lightbulbs to burn out or all the drugged rats to die. The lifetimes are observed in order, so let us denote the ordered observations by

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)}$$

and suppose that observations are stopped after the *sample α-quantile* is observed for some $0 < \alpha < 1$. This is the order statistic $X_{([n\alpha])}$, where $[n\alpha]$ is the greatest integer in $n\alpha$. The exponential distributions form a very common model in life-testing, so it is useful to test fit to this family given only the censored data

$$X_{(1)} < X_{(2)} < \ldots < X_{([n\alpha])}.$$

Here is a challenge to the great flexibility of chi-square methods. The response is to use random cells with boundaries given by sample δ_i-quantiles $\xi_i = X_{([n\delta_i])}$ for

$$0 = \delta_0 < \delta_1 < \ldots < \delta_{M-1} = \alpha < \delta_M = 1,$$

so that the $n - [n\alpha]$ unobserved lifetimes fall in the rightmost cell. Of course, $\xi_0 = 0$ and $\xi_M = \infty$. These random cells fit the demands of Section 6, for the sample quantile ξ_i converges in probability to the population δ_i-quantile $x_i(\theta)$ as $n \to \infty$. The population quantiles are found from

$$\int_0^{x_i} \frac{1}{\theta} e^{-x/\theta} dx = \delta_i.$$

This choice of cells produces *nonrandom* cell frequencies

$$N_i = [n\delta_i] - [n\delta_{i-1}],$$

but the theory of Section 6 is entirely undisturbed by this rather odd happenstance. We may cheerfully compute a variety of chi-square statistics for these cells, but we will content ourselves with the Pearson statistic. The grouped data MLE is computed exactly as in (27), "ignoring" the fact that the cell boundaries are random. That is, find $\bar{\theta}_n = \bar{\theta}_n(\xi_1, \ldots, \xi_{M-1})$ numerically by solving

$$\sum_{i=1}^{M} N_i \frac{\xi_{i-1} e^{-\xi_{i-1}/\theta} - \xi_i e^{-\xi_i/\theta}}{e^{-\xi_{i-1}/\theta} - e^{-\xi_i/\theta}} = 0,$$

then use the statistic

$$\sum_{i=1}^{M} \frac{[N_i - np_i(\bar{\theta}_n)]^2}{np_i(\bar{\theta}_n)}$$

with $p_i(\theta) = \xi_{i-1} e^{-\xi_{i-1}/\theta} - \xi_i e^{-\xi_i/\theta}$ and critical points from the $\chi^2(M-2)$ table. Hats off to the amazing chi-square statistics.

REFERENCES

1. Billingsley, Patrick (1968). *Convergence of Probability Measures.* Wiley, New York.

2. Chernoff, H. and Lehmann, E. L. (1954). The use of maximum-likelihood estimates in χ^2 tests for goodness of fit. *Annals of Mathematical Statistics* 25, 579-586.

3. Cramér, Harold (1946). *Mathematical Methods of Statistics.* Princeton University Press, Princeton, NJ.

4. Moore, David S. and Spruill, M. C. (1975). Unified large-sample theory of general chi-squared statistics for tests of fit. *Annals of Statistics* 3, 599-616.

5. Rao, K. C. and Robson, D. S. (1974). A chi-squared statistic for goodness-of-fit tests within the exponential family. *Communications in Statistics* 3, 1139-1153.

(19) (18) AFOSR TR-76-1227

CHI-SQUARE TESTS. (6)

(9) Tech. rept.

(14) mimeograph Series 469

7. AUTHOR(s)
David S. Moore (10)

(15) AFOSR-2354-72

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Purdue University
Department of Statistics
West Lafayette, IN 47907

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
61102F 2304/A5 (16) (17)

11. CONTROLLING OFFICE NAME AND ADDRESS
Air Force Office of Scientific Research/NM
Bolling AFB, Washington, DC 20332

12. REPORT DATE (11) 1976

13. NUMBER OF PAGES
49

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

(12) 52p.

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Goodness of fit
Chi-square tests
Quadratic forms

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
An exposition of chi-square tests in the context of mdern large-sample theory. The paper provides some historical and practical motivation and necessary background on large-sample methods and especially on multivariate normal random variables and their quadratic forms. There follows a unified exposition of the distribution of general chi-square statistics and of construction of statistics having a specified and limiting null distribution.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED 291 730
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)